DD2434 – Advanced Machine Learning

Lecture 3: **Linear regression, model comparison**

Pawel Herman

Computational Science and Technology (CST)

KTH Royal Institute of Technology

November 2018

Short outline for today

1. Recap from Lecture 2
2. Regression models
   a) maximum likelihood, regularization
   b) Bayesian approach
3. Model selection.

Main points in lecture 2

- Motivation for a probabilistic perspective of machine learning
- Other flavours of ML theory (e.g. PAC)
- General philosophy of a Bayesian approach

$$p(\mathbf{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathbf{w})\, p(\mathbf{w})}{p(\mathcal{D})}$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Main points in lecture 2

- Motivation for a probabilistic perspective of machine learning – scientific discipline
- General philosophy of a Bayesian approach
- Bayesian vs frequentist viewpoint

Main points in lecture 2

- Motivation for a probabilistic perspective of machine learning – scientific discipline
- General philosophy of a Bayesian approach
- Bayesian vs frequentist viewpoint
- Learning as inference, regression example

Learning as inference – recap

1) **Maximise the likelihood**

$$\mathcal{D} \rightarrow \boldsymbol{w}_{\mathrm{ML}} \qquad \ln p(\boldsymbol{t} \mid \mathbf{X}, \boldsymbol{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

2) Parameters of the model as random variables with a prior distribution

$$\mathcal{D}, p(\boldsymbol{w}) \rightarrow \boldsymbol{w}_{\mathrm{MAP}} \qquad p(\boldsymbol{w} \mid \alpha) = \mathcal{N}(\boldsymbol{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I})$$

Maximise the posterior: $\max\, p(\boldsymbol{w} \mid \mathbf{X}, \boldsymbol{t}, \alpha, \beta) \propto p(\boldsymbol{t} \mid \mathbf{X}, \boldsymbol{w}, \beta)\; p(\boldsymbol{w} \mid \alpha)$

3) **Marginalise** over parameters

Bayes: $\mathcal{D}, p(\boldsymbol{w}) \rightarrow p(\boldsymbol{w} \mid \mathcal{D})$

$$p(t \mid x, \mathbf{X}, \boldsymbol{t}) = \int p(t \mid x, \boldsymbol{w}, \beta)\; p(\boldsymbol{w} \mid \mathbf{X}, \boldsymbol{t}, \alpha, \beta)\, \mathrm{d}\boldsymbol{w}$$

Learning as inference – recap

1) $\mathcal{D} \rightarrow \boldsymbol{w}_{\mathrm{ML}}$

2) $\mathcal{D}, p(\boldsymbol{w}) \rightarrow \boldsymbol{w}_{\mathrm{MAP}}$

3) Bayes: $\mathcal{D}, p(\boldsymbol{w}) \rightarrow p(\boldsymbol{w} \mid \mathcal{D})$

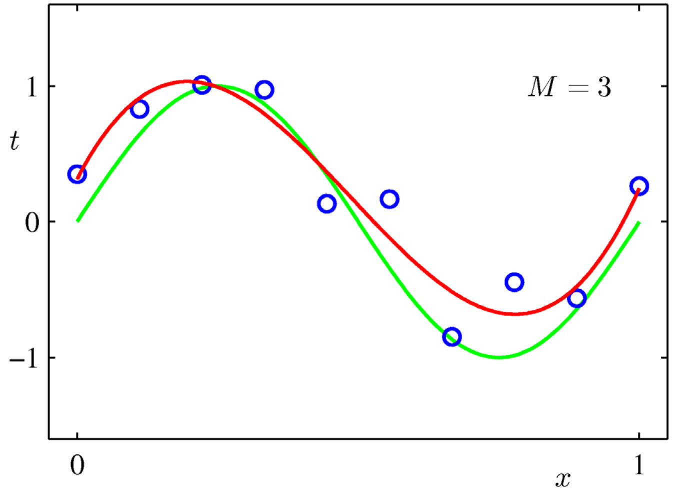
We will elaborate more particularly on the Bayesian regression in this lecture.

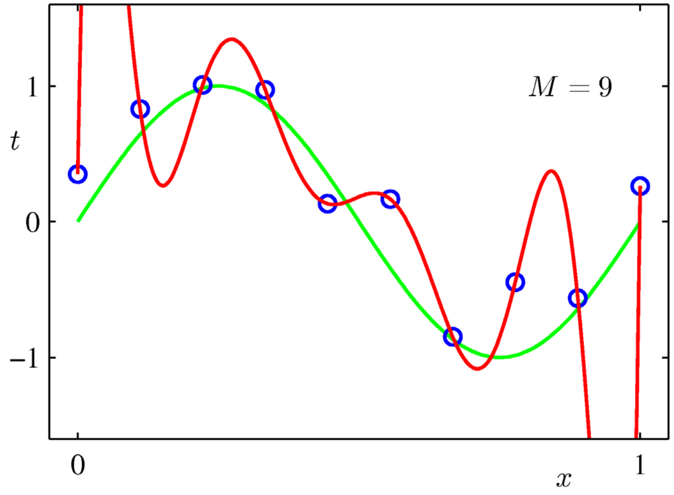
Main points in lecture 2

- Motivation for a probabilistic perspective of machine learning – scientific discipline
- General philosophy of a Bayesian approach
- Bayesian vs frequentist viewpoint
- Learning as inference, regression example
- Generative vs discriminative models
- Model complexity, overfitting, model selection

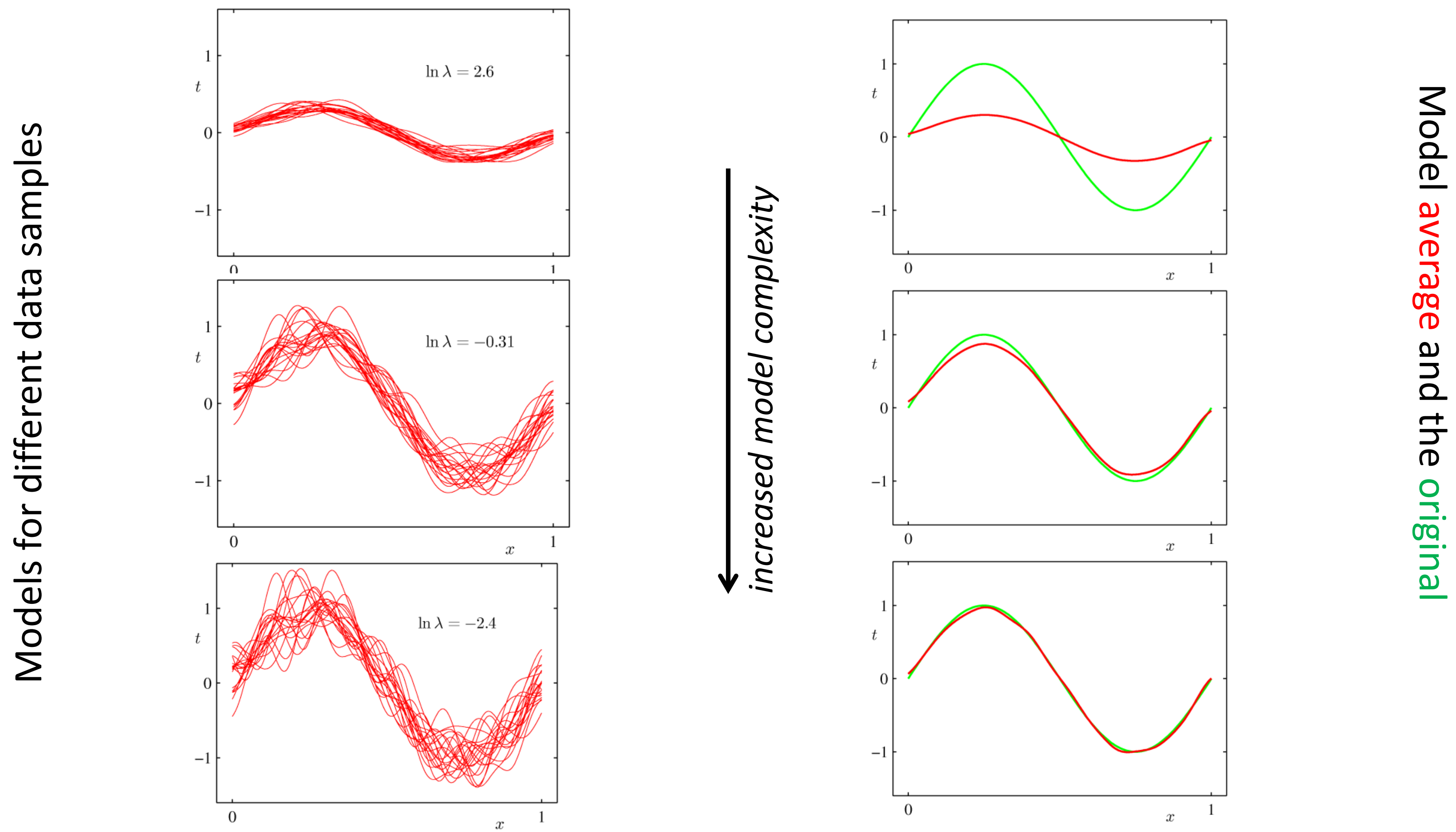
Model complexity

- Overfitting of maximum likelihood models



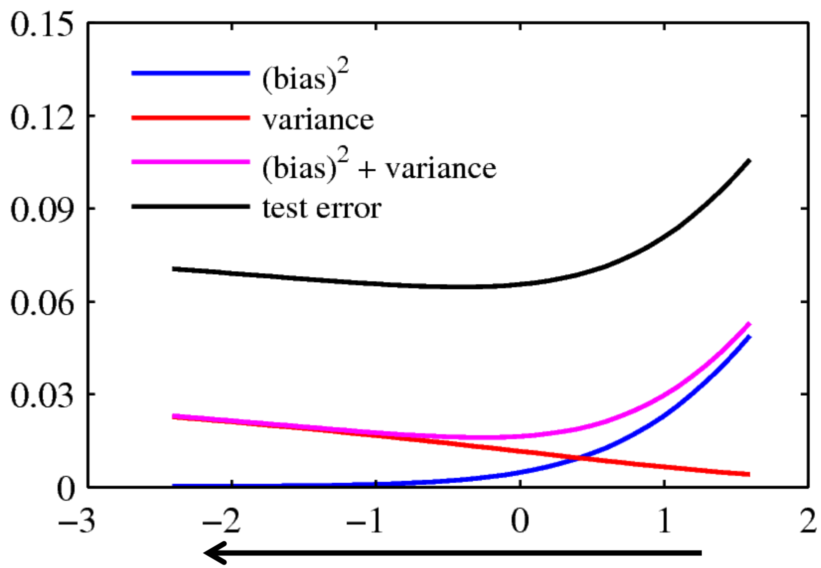


Bias-variance dilemma: model complexity



Bias-variance dilemma: model complexity

$$E[L] = (\text{bias})^2 + \text{variance} + \text{noise}$$



increased model complexity

Model selection

- Occam's razor – *"Accept the simplest explanation that fits the data."*
- Frequentist approach with maximum likelihood
  - bias-variance dilemma
  - need to control the model's complexity
    - regularisation
    - correction for the bias of ML estimates (AIC, BIC)
    - empirical estimate of generalisation error on a hold-out set (validation, resampling)
    - structural risk minimization (SRM) (minimise upper bound on the true risk), see also VC dimension (statistical learning theory)

Short outline for today

1. Recap from Lecture 2
2. Regression models
    a) maximum likelihood, regularization
    b) Bayesian approach, model selection
3. Model selection.

Linear regression – the "*work horse*" of ML

- Linear basis function models

$$y(\mathbf{x},\mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x})$$

$$\mathbf{w} = \left(w_0, \ldots, w_{M-1}\right)^{\mathrm{T}} \qquad \boldsymbol{\phi} = \left(\phi_0, \ldots, \phi_{M-1}\right)^{\mathrm{T}}$$

e.g., for polynomial one-dimensional regression basis functions are $\phi_j(x) = x^j$

Linear regression – the "*work horse*" of ML

- Linear basis function models

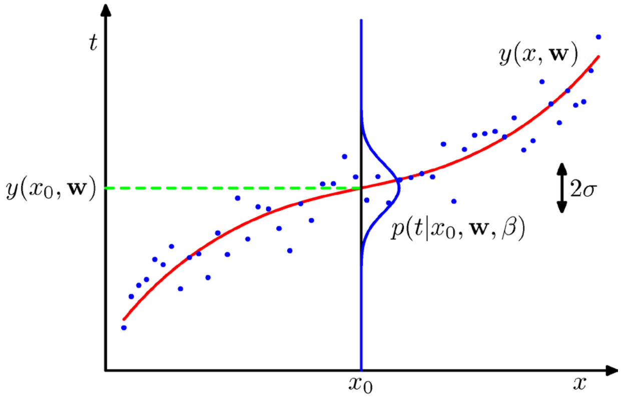
$$y(\mathbf{x},\mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x})$$

$$\mathbf{w} = \left(w_0,\ldots,w_{M-1}\right)^{\mathrm{T}} \qquad \boldsymbol{\phi} = \left(\phi_0,\ldots,\phi_{M-1}\right)^{\mathrm{T}}$$

e.g., for polynomial one-dimensional regression basis functions are $\phi_j(x) = x^j$

$$t = y(\mathbf{x},\mathbf{w}) + \varepsilon = \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}) + \varepsilon$$

$$p(t \mid \mathbf{x},\mathbf{w},\beta) = \mathcal{N}(t \mid y(\mathbf{x},\mathbf{w}), \beta^{-1})$$



Linear regression

- Maximum likelihood estimation

$$y(\mathbf{x},\mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x})$$

Likelihood:

$$p(\mathbf{t}_{\mathcal{D}_{trn}} \mid \mathbf{X}_{\mathcal{D}_{trn}}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}_n), \beta^{-1})$$

$$\mathbf{X}_{\mathcal{D}_{trn}} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$$

$$\mathbf{t}_{\mathcal{D}_{trn}} = (t_1, \ldots, t_N)$$

$$\Downarrow$$

$$\ln p(\mathbf{t}_{\mathcal{D}_{trn}} \mid \mathbf{X}_{\mathcal{D}_{trn}}, \mathbf{w}, \beta) = \underbrace{-\frac{\beta}{2} \sum_{n=1}^{N} \left\{ t_n - \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \right\}^2}_{E_D(\mathbf{w})} + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

Linear regression

- Maximum likelihood

$$0 = \sum_{n=1}^{N} t_n \boldsymbol{\phi}(\mathbf{x}_n) - \mathbf{w}^{\mathrm{T}} \left(\sum_{n=1}^{N} \boldsymbol{\phi}(\mathbf{x}_n) \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} \right)$$

Linear regression

- Maximum likelihood

$$0 = \sum_{n=1}^{N} t_n \boldsymbol{\phi}(\mathbf{x}_n) - \mathbf{w}^{\mathrm{T}} \left(\sum_{n=1}^{N} \boldsymbol{\phi}(\mathbf{x}_n) \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} \right)$$

Least-square solution:
(normal equations)

$$\mathbf{w}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t}$$

pseudo-inverse of the design matrix $\boldsymbol{\Phi}$

$$\beta_{\mathrm{ML}}^{-1} = \frac{1}{N} \sum_{n=1}^{N} \left\{ t_n - \mathbf{w}_{\mathrm{ML}}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \right\}^2$$

Linear regression

- Maximum likelihood

$$0 = \sum_{n=1}^{N} t_n \boldsymbol{\phi}(\mathbf{x}_n) - \mathbf{w}^{\mathrm{T}} \left(\sum_{n=1}^{N} \boldsymbol{\phi}(\mathbf{x}_n) \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} \right)$$

Least-square solution:
(normal equations)

$$\mathbf{w}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t}$$

pseudo-inverse of
the design matrix $\boldsymbol{\Phi}$

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$

Linear regression

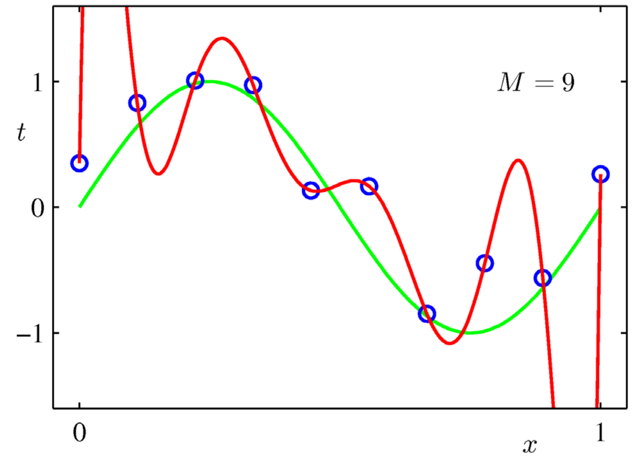
- Maximum likelihood

$$0 = \sum_{n=1}^{N} t_n \boldsymbol{\phi}(\mathbf{x}_n) - \mathbf{w}^{\mathrm{T}} \left(\sum_{n=1}^{N} \boldsymbol{\phi}(\mathbf{x}_n) \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} \right)$$

$$\begin{cases} \mathbf{w}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t} \\ \beta_{\mathrm{ML}}^{-1} = \dfrac{1}{N} \displaystyle\sum_{n=1}^{N} \left\{ t_n - \mathbf{w}_{\mathrm{ML}}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \right\}^2 \end{cases}$$

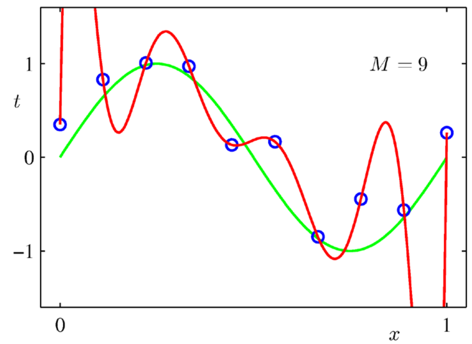
Linear regression - regularisation

- Problems with maximum likelihood estimate



Linear regression - regularisation

- Problems with maximum likelihood estimate



- We can address it by regularisation (parameter shrinkage)

$$\arg\min_{\mathbf{w}} \left\{ \frac{1}{2} \sum_{n=1}^{N} \{ t_n - \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \}^2 + \frac{\lambda}{2} \mathbf{w}^{\mathrm{T}} \mathbf{w} \right\}$$ (weight decay)

$$\arg\min_{\mathbf{w}} \left\{ \frac{1}{2} \sum_{n=1}^{N} \{ t_n - \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \}^2 + \frac{\lambda}{2} \sum_{j=1}^{M} |w_j|^q \right\}$$ (lasso for q=1)

Bayesian linear regression

- Bayesian treatment

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_0, \mathbf{S}_0)$$

Bayesian linear regression

- Bayesian treatment

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_0, \mathbf{S}_0)$$

For this **conjugate Gaussian prior** corresponding to the likelihood, the posterior has also Gaussian distribution:

$$p(\mathbf{w} \mid \mathbf{t}_{D_{trn}}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_N, \mathbf{S}_N)$$

$$\mathbf{m}_N = \mathbf{S}_N \left(\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \mathbf{\Phi}^{\mathrm{T}} \mathbf{t}_{\mathcal{D}_{trn}} \right)$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta \mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi}$$

$$\mathbf{\Phi}(\mathbf{X}_{\mathcal{D}_{trn}}) \to \mathbf{\Phi}$$

Bayesian linear regression

- Bayesian treatment

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I})$$

$$\Downarrow$$

$$p(\mathbf{w} \mid \mathbf{t}_{Dtrn}, \mathbf{X}_{\mathcal{D}_{trn}}, \alpha, \beta) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_N, \mathbf{S}_N)$$

$$\mathbf{m}_N = \beta \mathbf{S}_N \mathbf{\Phi}^{\mathrm{T}} \mathbf{t}_{\mathcal{D}_{trn}}$$

$$\mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi}$$

Bayesian linear regression

- Bayesian treatment

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I})$$

$$\Downarrow$$

the **likelihood part** of the log-posterior dependent on $\mathbf{w}$

the **prior part** of the log-posterior dependent on $\mathbf{w}$

$$\underbrace{\ln p(\mathbf{w} \mid \mathbf{t}_{\mathcal{D}_{trn}}, \mathbf{X}_{\mathcal{D}_{trn}}, \alpha, \beta)}_{\text{log-posterior}} = \boxed{-\frac{\beta}{2}\sum_{n=1}^{N}\left\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\}^2} \boxed{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}} + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

Bayesian linear regression

- Predictive distribution

$$\underbrace{p(t \mid \mathbf{x}, \mathbf{t}_{\mathcal{D}_{trn}}, \alpha, \beta)}_{\text{predictive}} = \int \underbrace{p(t \mid \mathbf{x}, \mathbf{w}, \beta)}_{\text{“noise” model}} \; \underbrace{p(\mathbf{w} \mid \mathbf{t}_{\mathcal{D}_{trn}}, \mathbf{X}_{\mathcal{D}_{trn}}, \alpha, \beta)}_{\text{posterior}} \, \mathrm{d}\mathbf{w}$$

Bayesian linear regression

- Predictive distribution

$$\underbrace{p(t \mid \mathbf{x}, \mathbf{t}_{\mathcal{D}_{trn}}, \alpha, \beta)}_{\text{predictive}} = \int \underbrace{p(t \mid \mathbf{x}, \mathbf{w}, \beta)}_{\text{"noise" model}} \; \underbrace{p(\mathbf{w} \mid \mathbf{t}_{\mathcal{D}_{trn}}, \mathbf{X}_{\mathcal{D}_{trn}}, \alpha, \beta)}_{\text{posterior}} \, \mathrm{d}\mathbf{w}$$

$$p(t \mid \mathbf{x}, \mathbf{t}_{Dtrn}, \alpha, \beta) = \mathcal{N}(t \mid \mathbf{m}_N^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x}))$$

$$\mathbf{m}_N = \beta \mathbf{S}_N \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t}_{\mathcal{D}_{trn}}$$

$$\mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}$$

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})$$

Bayesian linear regression

- Bayesian philosophy – overview
  - in order to make predictions we need posterior

$$p(\mathbf{w} \,|\, \mathbf{t}, \mathbf{x})$$

  - it can be updated (estimated) after the relevant information has been taken into account

Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Bayesian linear regression

- Bayesian philosophy
  - in order to make predictions we need posterior

$$p(\mathbf{w} \mid \mathbf{t}, \mathbf{x})$$

  - it can be updated (estimated) after the relevant information has been taken into account
  - so, we need our belief about the model and the observations (data: $\mathbf{t}$, $\mathbf{x}$)

Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Bayesian linear regression

- Bayesian philosophy
  - in order to make predictions we need posterior

$$p(\mathbf{w}\,|\,\mathbf{t}, \mathbf{x})$$

  - it can be updated (estimated) after the relevant information has been taken into account
  - so, we need our belief about the model and the observations (data: $\mathbf{t}$, $\mathbf{x}$)
  - before we see the data, we express our belief about the regression params

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$$

Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Bayesian linear regression

- Bayesian philosophy
  - in order to make predictions we need posterior

$$p(\mathbf{w}\,|\,\mathbf{t},\,\mathbf{x})$$

  - it can be updated (estimated) after the relevant information has been taken into account
  - so, we need our belief about the model and the observations (data: $\mathbf{t}$, $\mathbf{x}$)
  - before we see the data, we express our belief about the regression params

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$$

very often we do not really know too much in advance

Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Bayesian linear regression

- Bayesian philosophy (cont.)
  - how well does my model predicts (explains) data?
    - likelihood (like an error function) for a single sample n

$$p(t_n \mid \mathbf{w}, \mathbf{x}_n) = \mathcal{N}(t_n \mid \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\mathbf{I})$$

    - if there is independence between samples, likelihood for all data $\mathbf{t}$ amounts to

$$p(\mathbf{t} \mid \mathbf{w}, \mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\mathbf{I})$$

Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Bayesian linear regression

- Bayesian philosophy – summary
  - The goal is to reach posterior distribution after <u>all relevant</u> information has been taken into account.
  - Prediction should reflect my beliefs in the model and the information in the observations.

Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Bayesian linear regression

- Bayesian philosophy – summary
  - The goal is to reach posterior distribution after <u>all relevant</u> information has been taken into account.
  - Prediction should reflect my beliefs in the model and the information in the observations.
  - So:
    1. Choose a model
    2. Formulate prediction error by likelihood
    3. Formulate belief of a model in prior
    4. Marginalise irrelevant variables (parameters)

Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Bayesian linear regression

- Bayesian philosophy – summary
  - The goal is to reach posterior distribution after <u>all relevant</u> information has been taken into account.
  - Prediction should reflect my beliefs in the model and the information in the observations.
  - So:
    1. Choose a model
    2. Formulate prediction error by likelihood
    3. Formulate belief of a model in prior
    4. Marginalise irrelevant variables (parameters)

Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Bayesian linear regression

- Bayesian philosophy – summary
  - The goal is to reach posterior distribution after <u>all relevant</u> information has been taken into account.
  - Prediction should reflect my beliefs in the model and the information in the observations.
  - So:

Gaussian distributions are *self-conjugate*, i.e.:

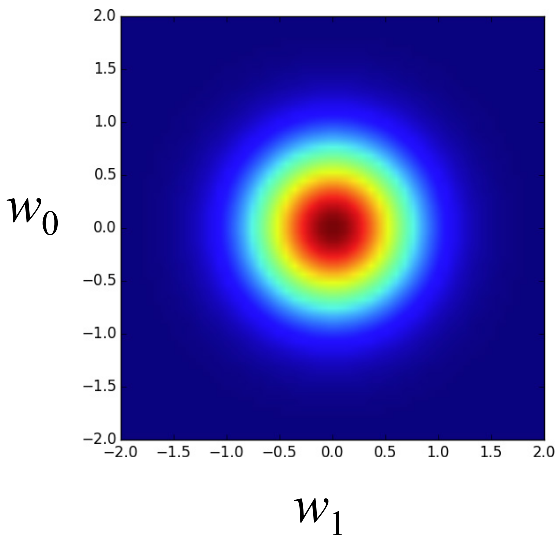
Gaussian prior + Gaussian likelihood → **Gaussian posterior**

Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Sequential Bayesian learning

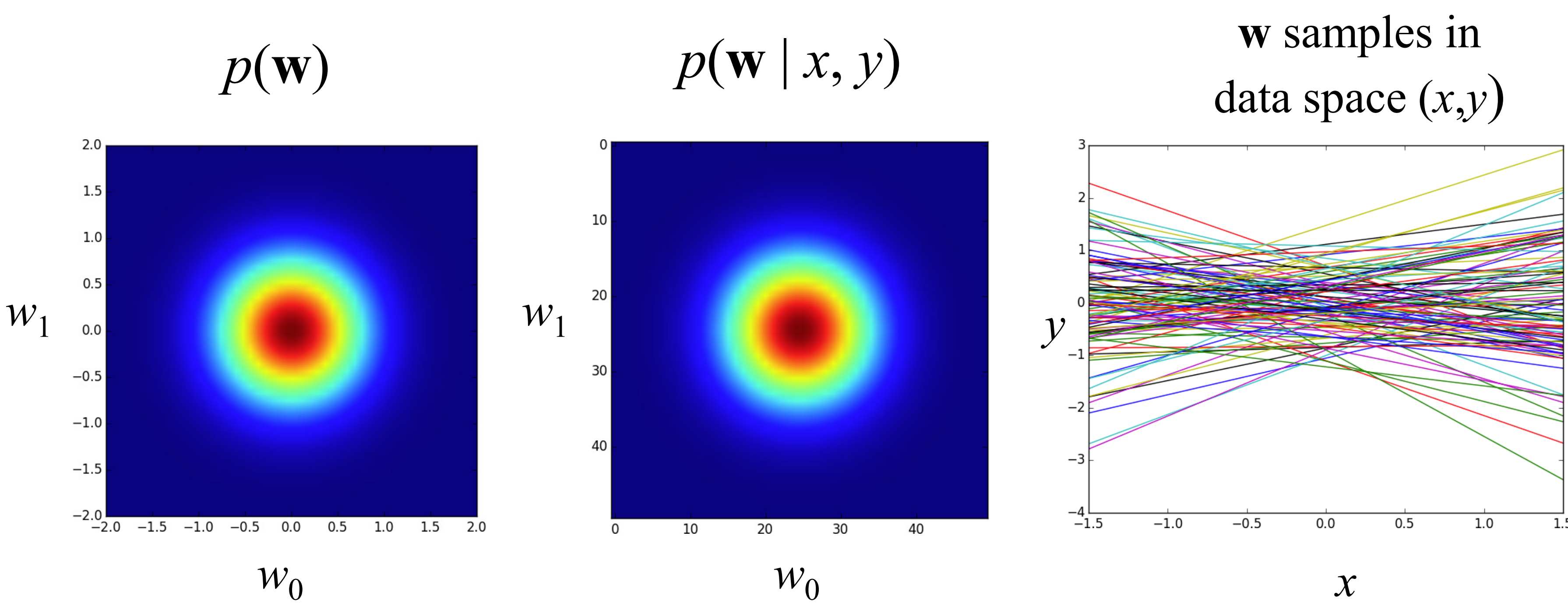
$$y(x,\mathbf{w}) = -w_0 + w_1 x$$

$$p(\mathbf{w})$$



Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

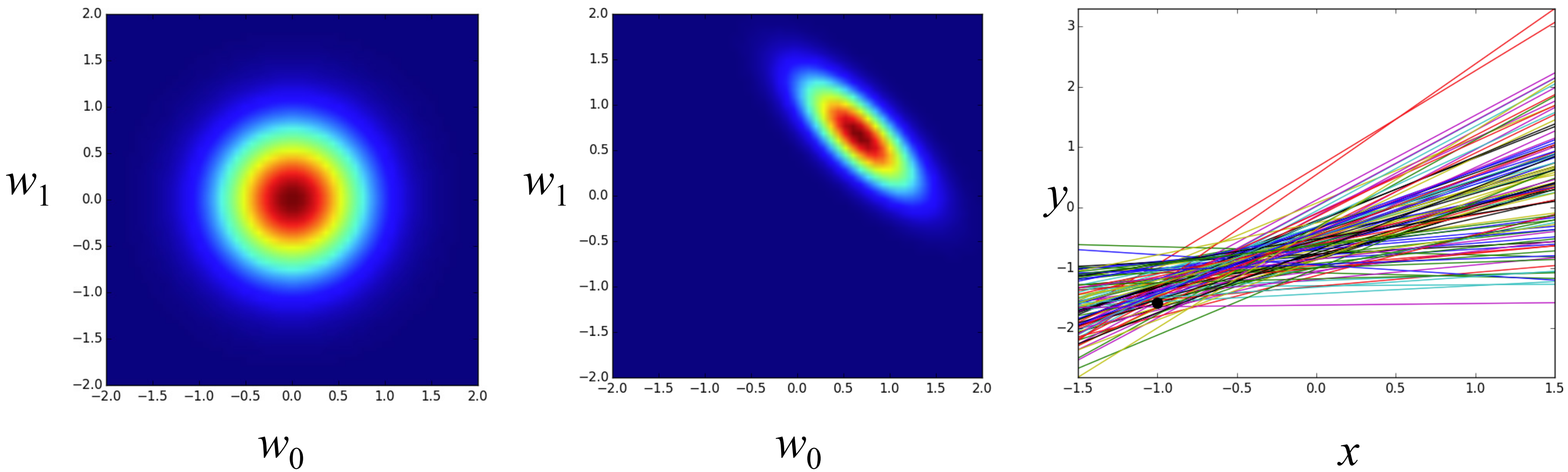
Sequential Bayesian learning



Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Sequential Bayesian learning

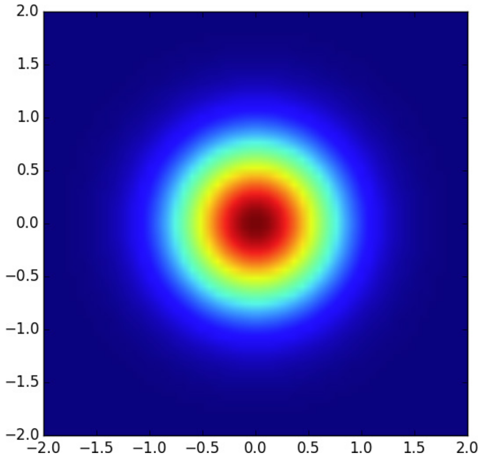
$p(\mathbf{w})$ $\quad$ $p(\mathbf{w} \mid x, y)$ $\quad$ $\mathbf{w}$ samples in data space (x,y)



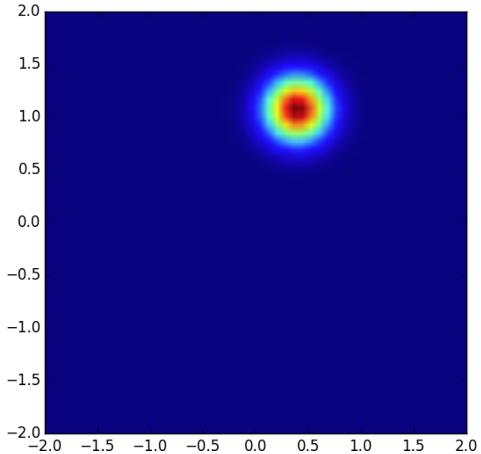
Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Sequential Bayesian learning

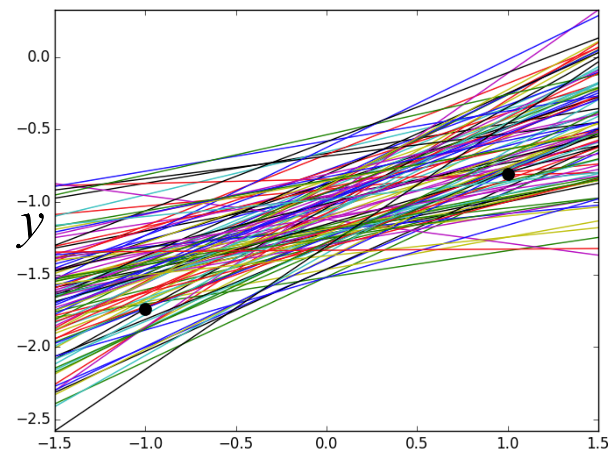
$p(\mathbf{w})$



$p(\mathbf{w} \mid x, y)$



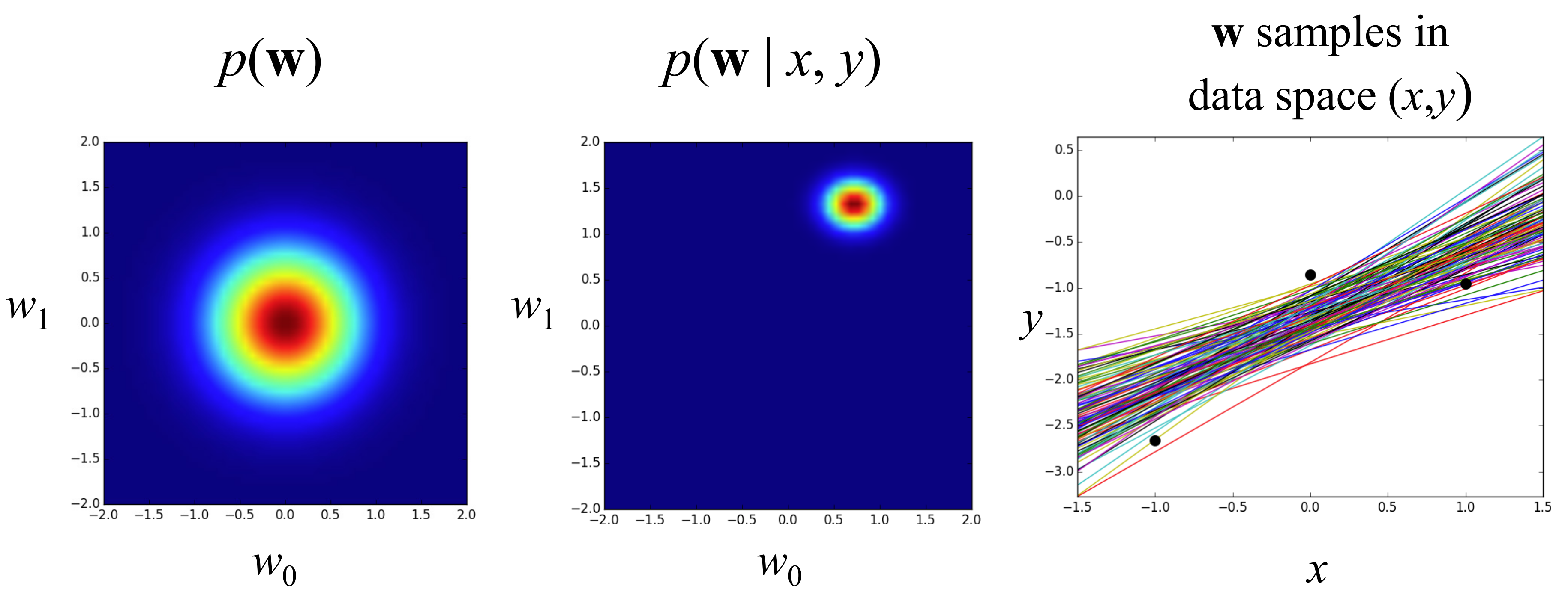
$\mathbf{w}$ samples in data space (x,y)



w_1 w_0 w_1 w_0 y x

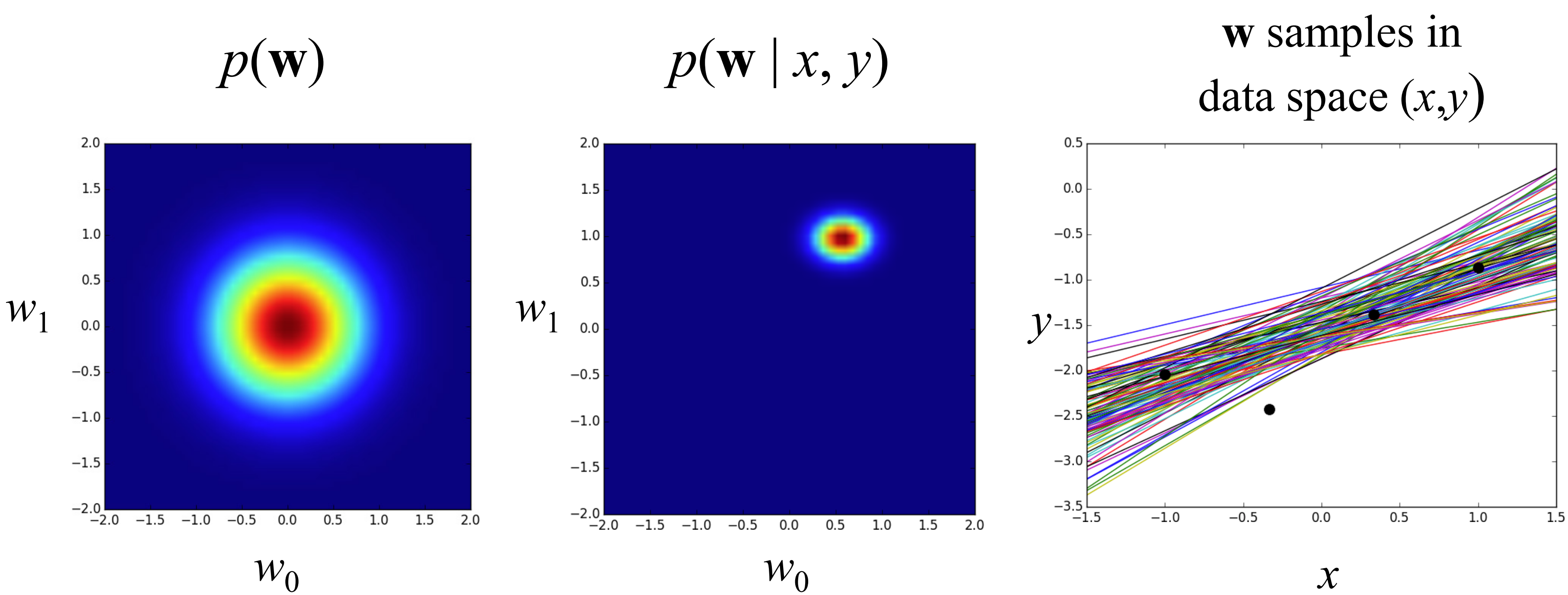
Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Sequential Bayesian learning



Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

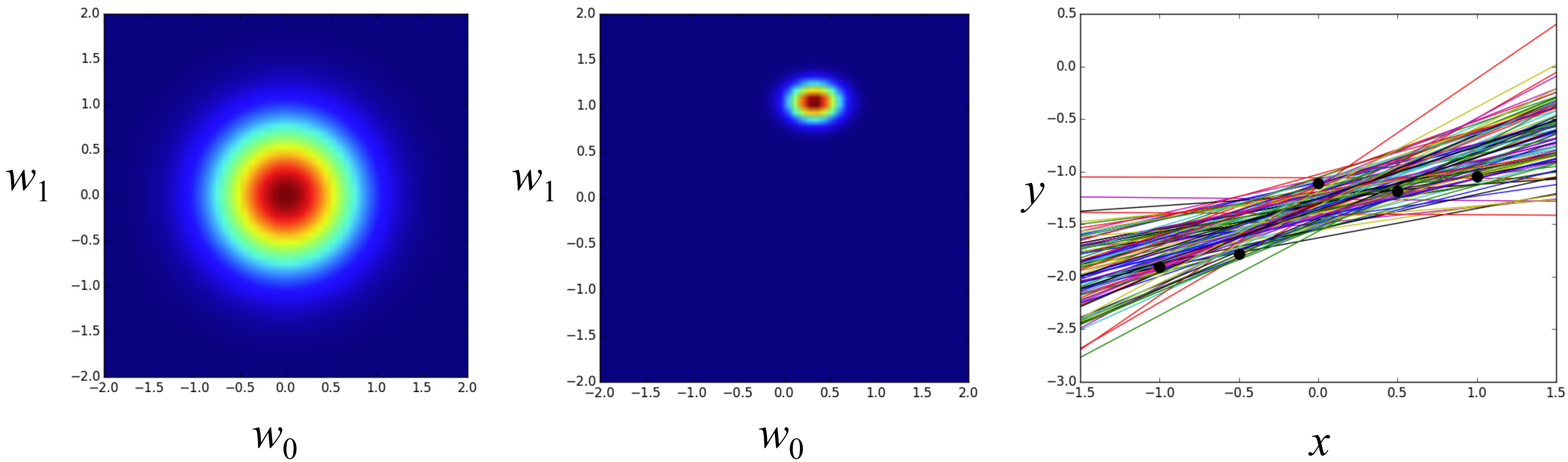
Sequential Bayesian learning



Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

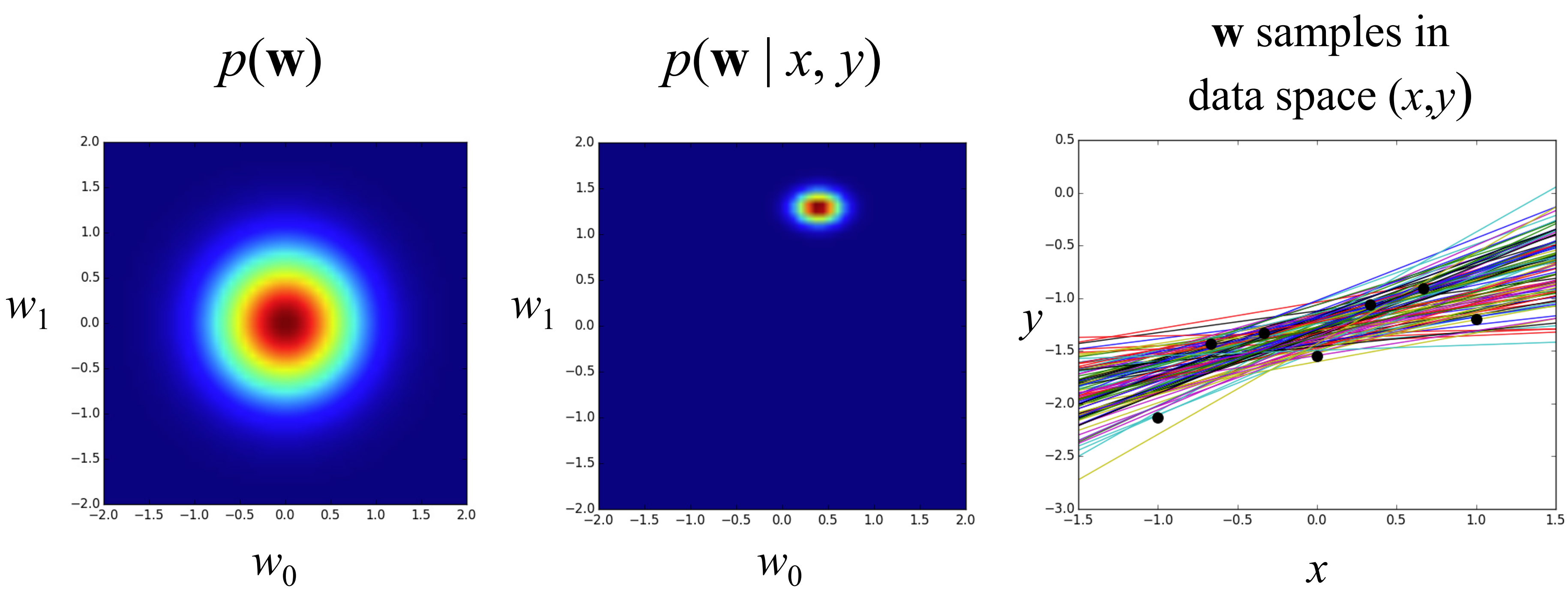
Sequential Bayesian learning

$p(\mathbf{w})$ $\qquad$ $p(\mathbf{w} \mid x, y)$ $\qquad$ $\mathbf{w}$ samples in data space (x,y)



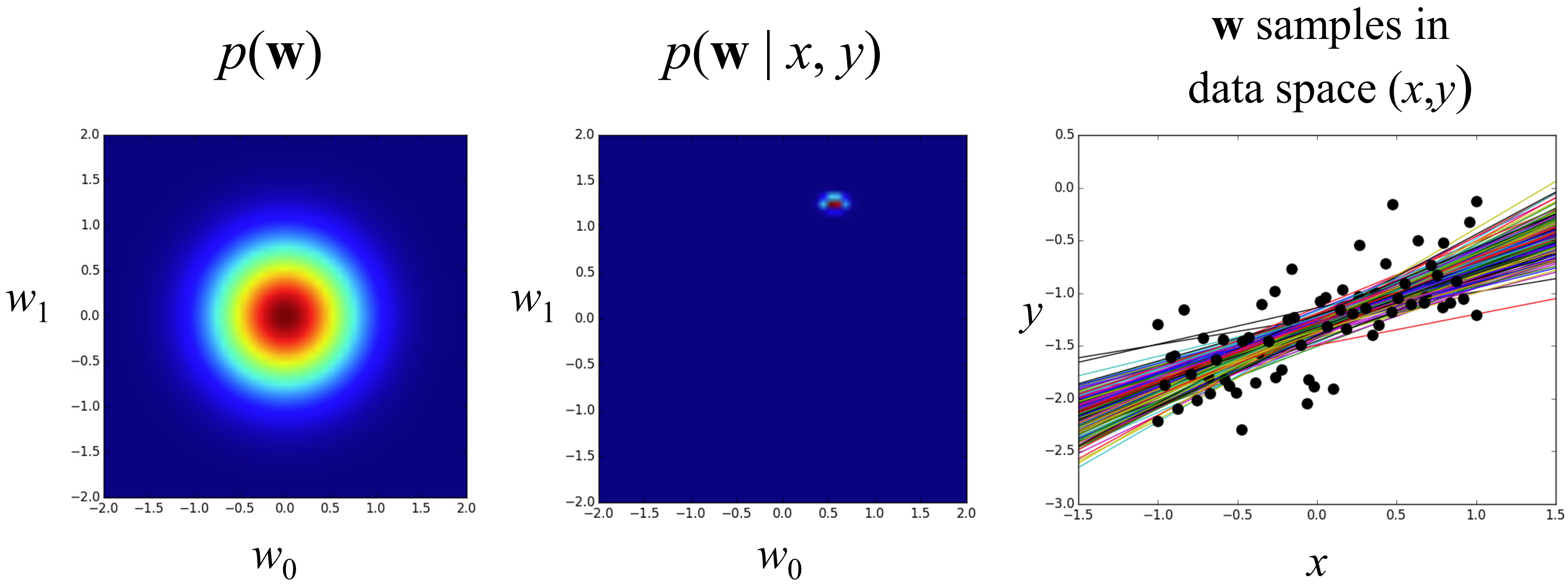
Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Sequential Bayesian learning



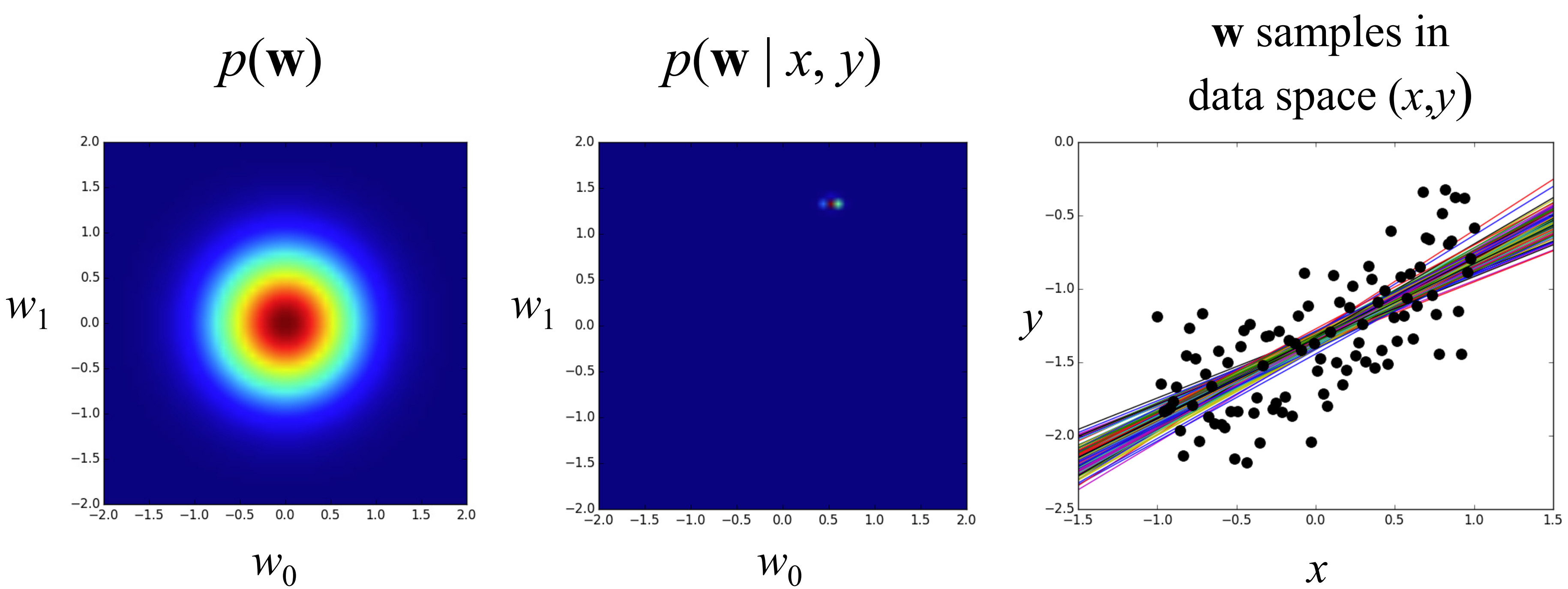
Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Sequential Bayesian learning



Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Sequential Bayesian learning



Adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Short outline for today

1. Recap from Lecture 2
2. Regression models
   a) maximum likelihood, regularization
   b) Bayesian approach, model selection
3. Model selection.

Bayesian model comparison

- Evidence framework

$$p(\mathbf{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathbf{w})\, p(\mathbf{w})}{p(\mathcal{D})}$$

  - The denominator does not change with $\mathbf{w}$

$$p(\mathbf{w} \mid \mathbf{t}_{\mathcal{D}}, \mathbf{X}_{\mathcal{D}}) \propto p(\mathbf{t}_{\mathcal{D}} \mid \mathbf{w}, \mathbf{X}_{\mathcal{D}})\, p(\mathbf{w})$$

Bayesian model comparison

- Evidence framework

$$p(\mathbf{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathbf{w})\, p(\mathbf{w})}{p(\mathcal{D})}$$

  - The denominator does not change with $\boldsymbol{w}$

$$p(\mathbf{w} \mid \mathbf{t}_{\mathcal{D}}, \mathbf{X}_{\mathcal{D}}) \propto p(\mathbf{t}_{\mathcal{D}} \mid \mathbf{w}, \mathbf{X}_{\mathcal{D}})\, p(\mathbf{w})$$

  - $p(\mathcal{D})$ shows where the model spreads its probability mass over the data space (evidence of the model)

$$p(\mathbf{w} \mid \mathcal{D}, \mathcal{M}_i) = \frac{p(\mathcal{D} \mid \mathbf{w}, \mathcal{M}_i)\, p(\mathbf{w} \mid \mathcal{M}_i)}{p(\mathcal{D} \mid \mathcal{M}_i)}$$

Bayesian model comparison

- What can we do with model evidence?

posterior: $p(\mathcal{M}_i \mid \mathcal{D}) \propto p(\mathcal{M}_i)\, p(\mathcal{D} \mid \mathcal{M}_i)$

marginal likelihood

Bayesian model comparison

- What can we do with model evidence?

posterior: $p(\mathcal{M}_i \mid \mathcal{D}) \propto p(\mathcal{M}_i)\, p(\mathcal{D} \mid \mathcal{M}_i)$

  - for the same priors, the posterior ratio between $\mathcal{M}_1$ and $\mathcal{M}_2$

$$\frac{p(\mathcal{D} \mid \mathcal{M}_1)}{p(\mathcal{D} \mid \mathcal{M}_2)}$$ Bayes factor

Bayesian model comparison

- What can we do with model evidence?

posterior: $p(\mathcal{M}_i \mid \mathcal{D}) \propto p(\mathcal{M}_i)\, p(\mathcal{D} \mid \mathcal{M}_i)$

  - for the same priors, the posterior ratio between $\mathcal{M}_1$ and $\mathcal{M}_2$

$$\frac{p(\mathcal{D} \mid \mathcal{M}_1)}{p(\mathcal{D} \mid \mathcal{M}_2)} \quad \text{Bayes factor}$$

  - how do we find the evidence $p(\mathcal{D} \mid \mathcal{M}_i)$?

$$p(\mathcal{D} \mid \mathcal{M}_i) = \int p(\mathcal{D} \mid \mathbf{w}, \mathcal{M}_i)\, p(\mathbf{w} \mid \mathcal{M}_i)\, d\mathbf{w}$$

Bayesian model comparison

- What can we do with model evidence?

posterior: $$p(\mathcal{M}_i \mid \mathcal{D}) \propto p(\mathcal{M}_i) p(\mathcal{D} \mid \mathcal{M}_i)$$

- for the same priors, the posterior ratio between $\mathcal{M}_1$ and $\mathcal{M}_2$

$$\frac{p(\mathcal{D} \mid \mathcal{M}_1)}{p(\mathcal{D} \mid \mathcal{M}_2)}$$ Bayes factor

- how do we find the evidence $p(\mathcal{D} \mid \mathcal{M}_i)$?

$$p(\mathcal{D} \mid \mathcal{M}_i) = \int p(\mathcal{D} \mid \mathbf{w}, \mathcal{M}_i) p(\mathbf{w} \mid \mathcal{M}_i) d\mathbf{w}$$

"The evidence can be seen as the probability of generating the data set from a model whose parameters are sampled at random from the prior" *Bishop, sec.3.4*

Bayesian model comparison

- What can we do with model evidence?

posterior: $p(\mathcal{M}_i \mid \mathcal{D}) \propto p(\mathcal{M}_i)\, p(\mathcal{D} \mid \mathcal{M}_i)$

  - prediction can be made using the “best” model (single most probable model) or can be averaged from the mixture of K models

Model mixture: $$p(t \mid \mathbf{x}, \mathcal{D}) = \sum_{i=1}^{K} p(t \mid \mathbf{x}, \mathcal{M}_i, \mathcal{D})\, p(\mathcal{M}_i \mid \mathcal{D})$$

Bayesian model comparison

- Simple approximation for a single parameter w

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|w,\mathcal{M}_i)\,p(w|\mathcal{M}_i)\,dw$$

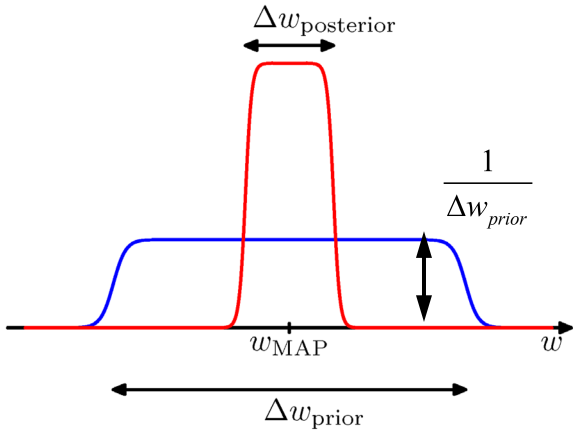
Bayesian model comparison

- Simple approximation for a single parameter w

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|w,\mathcal{M}_i)\,p(w|\mathcal{M}_i)\,dw$$

  - If the *posterior* is sharply peaked around some w_{MAP}, with the width $\Delta w_{\mathrm{posterior}}$, and the *prior* is flat with the width $\Delta w_{\mathrm{prior}}$, then

$$p(\mathcal{D}|\mathcal{M}_i) \approx p(\mathcal{D}|w_{\mathrm{MAP}},\mathcal{M}_i)\frac{\Delta w_{posterior}}{\Delta w_{prior}}$$



Bayesian model comparison

- Simple approximation for a single parameter w

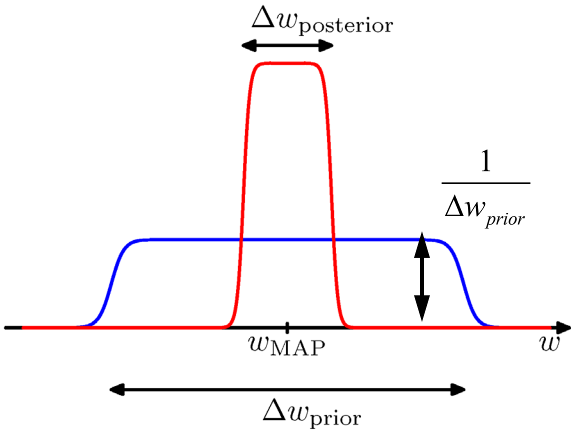
$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|w,\mathcal{M}_i)\,p(w|\mathcal{M}_i)\,dw$$

  - If the *posterior* is sharply peaked around some w_{MAP}, with the width $\Delta w_{\mathrm{posterior}}$, and the *prior* is flat with the width $\Delta w_{\mathrm{prior}}$, then

$$p(\mathcal{D}|\mathcal{M}_i) \approx p(\mathcal{D}|w_{\mathrm{MAP}},\mathcal{M}_i)\frac{\Delta w_{posterior}}{\Delta w_{prior}}$$

$$\Downarrow$$

$$\ln p(\mathcal{D}|\mathcal{M}_i) \approx \ln p(\mathcal{D}|w_{\mathrm{MAP}},\mathcal{M}_i) + \ln\left(\frac{\Delta w_{posterior}}{\Delta w_{prior}}\right)$$



Bayesian model comparison

- Simple approximation for a single parameter w

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|w,\mathcal{M}_i)\,p(w|\mathcal{M}_i)\,dw$$

  - If the posterior is sharply peaked around some w_{MAP}, with the width $\Delta w_{\mathrm{posterior}}$, and the prior is flat with the width $\Delta w_{\mathrm{prior}}$, then

$$p(\mathcal{D}|\mathcal{M}_i) \approx p(\mathcal{D}|w_{\mathrm{MAP}},\mathcal{M}_i)\frac{\Delta w_{posterior}}{\Delta w_{prior}}$$

$$\Downarrow$$

$$\ln p(\mathcal{D}|\mathcal{M}_i) \approx \ln p(\mathcal{D}|w_{\mathrm{MAP}},\mathcal{M}_i) + M \ln\left(\frac{\Delta w_{posterior}}{\Delta w_{prior}}\right)$$

$\Bigg\{$ Similar estimate can be made for M parameters (assuming their distributions behave similarly)

Bayesian model comparison

- Simple approximation for a single parameter w

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|w,\mathcal{M}_i)\,p(w|\mathcal{M}_i)\,dw$$

  - If the posterior is sharply peaked around some w_{MAP}, with the width $\Delta w_{\text{posterior}}$, and the prior is flat with the width Δw_{prior}, then

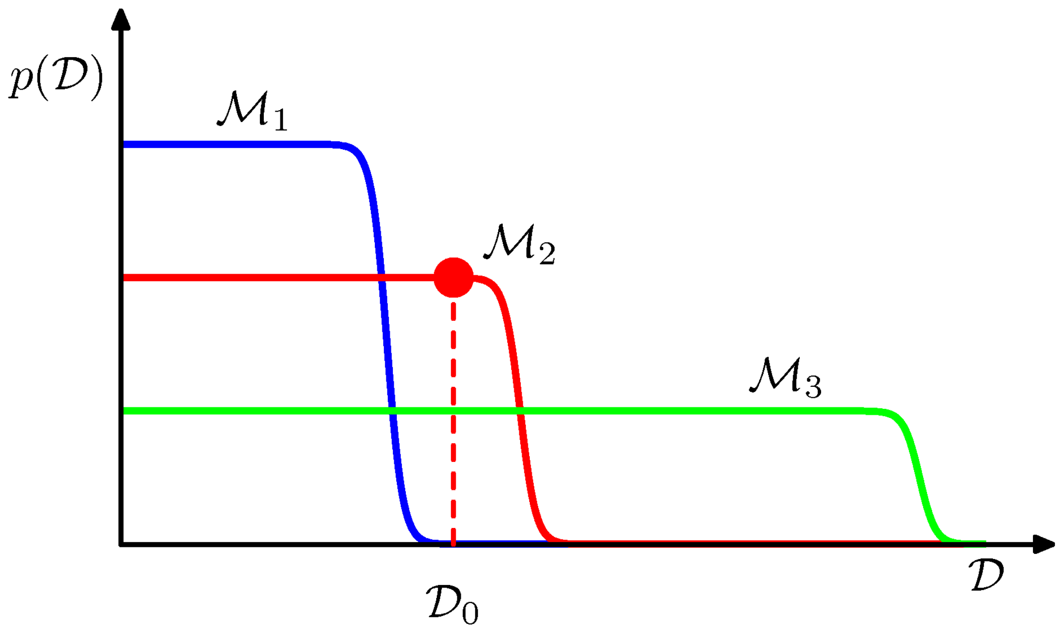
$$p(\mathcal{D}|\mathcal{M}_i) \approx p(\mathcal{D}|w_{\text{MAP}},\mathcal{M}_i)\frac{\Delta w_{posterior}}{\Delta w_{prior}}$$

$$\Downarrow$$

$$\ln p(\mathcal{D}|\mathcal{M}_i) \approx \ln p(\mathcal{D}|w_{\text{MAP}},\mathcal{M}_i) + {\color{red}M}\ln\left(\frac{\Delta w_{posterior}}{\Delta w_{prior}}\right)$$

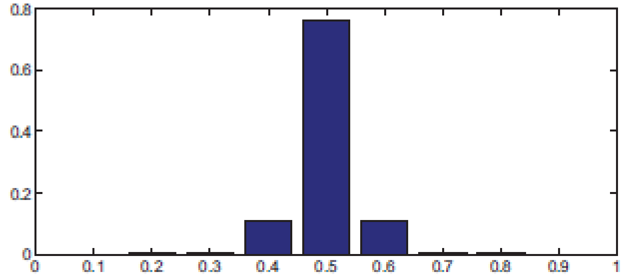
It can be seen as *Occam factor.*

Bayesian model comparison

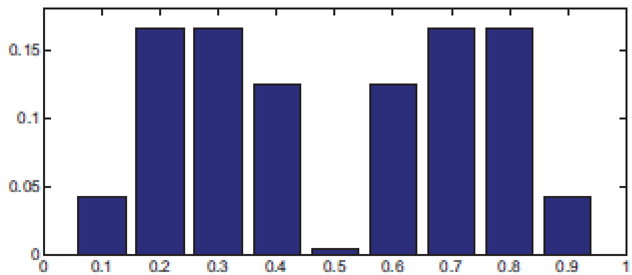


Example – a discrete parameter space

We have two discrete models: "fair" coin (M_1) and "biased" coin (M_2) model



prior for θ for "fair" coin (M_1)



prior for θ for "biased" coin (M_2)

The evidence for each model is:

$$p(\mathcal{D}|\,\mathcal{M}_i) = \sum_{\theta} p(\mathcal{D}|\,\theta, \mathcal{M}_i)\, p(\theta\,|\,\mathcal{M}_i)$$

Example – a discrete parameter space

Evidence for M_1 and M_2

$$p(\mathcal{D}|\,\mathcal{M}_i) = \sum_{\theta} p(\mathcal{D}|\,\theta, \mathcal{M}_i)\, p(\theta\,|\,\mathcal{M}_i)$$

For N_H heads and N_T tails, we obtain the following evidence for model i:

$$p(\mathcal{D}|\,\mathcal{M}_i) = \sum_{\theta} \theta^{N_H} (1-\theta)^{N_T} \;\; p(\theta|\,\mathcal{M}_i)$$

Example – a discrete parameter space

Evidence for M_1 and M_2

$$p(\mathcal{D}|\mathcal{M}_i) = \sum_{\theta} p(\mathcal{D}|\theta,\mathcal{M}_i)\,p(\theta\,|\,\mathcal{M}_i)$$

For N_H heads and N_T tails, we obtain the following evidence for model i:

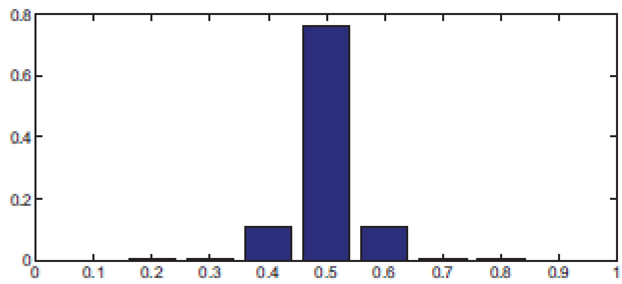
$$p(\mathcal{D}|\mathcal{M}_i) = \sum_{\theta} \theta^{N_H}(1-\theta)^{N_T}\;\; p(\theta|\mathcal{M}_i)$$

If we assume that $p(M_1)$=$p(M_2)$, i.e. both “fair” &“biased” coins are equally probable, then the Bayes’ factor is decisive:

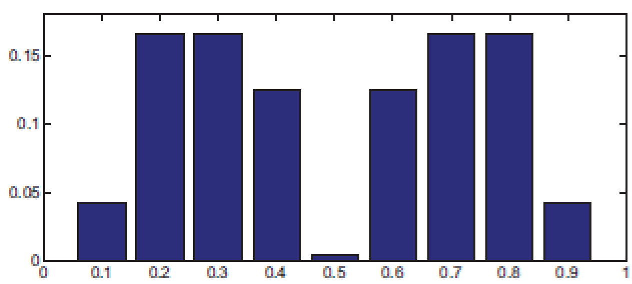
$$\frac{p(\mathcal{M}_{fair}\,|\,\mathcal{D})}{p(\mathcal{M}_{biased}\,|\,\mathcal{D})} = \frac{p(\mathcal{D}|\mathcal{M}_{fair})}{p(\mathcal{D}|\mathcal{M}_{biased})}$$

Example – a discrete parameter space

$$\theta \in \{0.1, 0.2, \ldots, 0.9\}$$



prior for θ for "fair" coin (M_1)



prior for θ for "biased" coin (M_2)

$$p(\mathcal{D} | \theta) = \theta^{N_H} (1-\theta)^{N_T}$$

Example – a discrete parameter space

$$\theta \in \{0.1, 0.2, \ldots, 0.9\}$$

Bayes' factor estimates:

N_H = 5 and N_T = 2

$$\frac{p(\mathcal{D} \mid \mathcal{M}_{fair})}{p(\mathcal{D} \mid \mathcal{M}_{biased})} = 1.09$$

N_H = 50 and N_T = 20

$$\frac{p(\mathcal{D} \mid \mathcal{M}_{fair})}{p(\mathcal{D} \mid \mathcal{M}_{biased})} = 0.109$$

The evidence approximation

If hyperpriors over α and β are introduced, we obtain the predictive distribution by marginalizing out $\mathbf{w}$, α and β

$$p(t \mid \mathbf{t}, \mathbf{x}) = \iiint p(t \mid \mathbf{w}, \mathbf{x}, \beta)\, p(\mathbf{w} \mid \mathbf{t}, \alpha, \beta)\, p(\alpha, \beta \mid \mathbf{t})\, dw\, d\alpha\, d\beta$$

The evidence approximation

If hyperpriors over α and β are introduced, we obtain the predictive distribution by marginalizing out $\mathbf{w}$, α and β

$$p(t \mid \mathbf{t}, \mathbf{x}) = \iiint p(t \mid \mathbf{w}, \mathbf{x}, \beta)\, p(\mathbf{w} \mid \mathbf{t}, \alpha, \beta)\, p(\alpha, \beta \mid \mathbf{t})\, dw\, d\alpha\, d\beta$$

From Bayes theorem:

$$p(\alpha, \beta \mid \mathbf{t}) \propto p(\mathbf{t} \mid \alpha, \beta)\, p(\alpha)\, p(\beta)$$

The evidence approximation

If hyperpriors over α and β are introduced, we obtain the predictive distribution by marginalizing out $\mathbf{w}$, α and β

$$p(t \mid \mathbf{t}, \mathbf{x}) = \iiint p(t \mid \mathbf{w}, \mathbf{x}, \beta)\, p(\mathbf{w} \mid \mathbf{t}, \alpha, \beta)\, p(\alpha, \beta \mid \mathbf{t})\, dw\, d\alpha\, d\beta$$

From Bayes theorem:

$$p(\alpha, \beta \mid \mathbf{t}) \propto p(\mathbf{t} \mid \alpha, \beta)\, p(\alpha)\, p(\beta)$$

For flat prior over α and β we can resort to maximising the ***marginal likelihood***.

(*type II maximum likelihood*)